1019-1002

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re the Application of: Fultheim, Shai	Application No.: 10/828,465
Filed: 04/21/04	Art Unit: 2128
For: Cluster Based Operating System-Agnostic Virtual Computing System	Examiner: David Silver

DECLARATION UNDER 37 CFR 1.132

I, the undersigned, Joseph Landman, hereby declare as follows:

1) I am making this Declaration in support of the patentability of the claims in U.S. Patent Application 10/828,465 (referred to hereinafter as "the Application"). Specifically, this Declaration will set forth my opinion, with supporting facts and evidence, that:

(a) The references that the Examiner has cited against the claims in the Application could not have led a person of ordinary skill in the art to make the claimed invention; and

(b) The invention defined by the claims in the Application, embodied in the vSMP product sold by ScaleMP (the assignee of the Application), answers a market need that could not be satisfied by prior art solutions and has been enthusiastically received by market leaders and customers.

2) I am not an employee of ScaleMP and have no economic interest in the company. I have not received compensation for my services in preparing this Declaration.

3) I have worked in the computer industry for more than 20 years, specializing in high-performance computing (HPC). I received my B.A. in physics in from the State University of New York at Stony Brook in 1987, my M.S. in physics from Michigan State University in 1990, and my Ph.D. in computational physics from Wayne State University in 1997. I have worked as an engineer., a researcher at a number of leading computer and software companies, including IBM and Silicon Graphics. I have also served as a part time research assistant professor in the computer science department at Wayne State University, teaching high performance computing programming. In 2002 I founded my own company, Scalable Informatics, which provides solutions for HPC users, including supercomputers, clusters, and storage systems. My detailed *curriculum vitae* is attached hereto as Exhibit A.

4) I am considered to be an industry authority in the field of HPC and have a detailed, first-hand understanding of both the technical and marketing aspects of this field. I publish a widely-read blog on HPC at scalability.org. I have given talks at various HPC and scientific venues over the last 12 years, including Linux World, ISMB, GSAC, and AINA. I have written white papers under contract to multiple organizations surrounding topics in HPC systems and applications and have published peer-reviewed papers in ACM and IEEE journals. I have also collaborated in development of a number of new HPC systems and programming tools, as detailed in my *curriculum vitae*.

5) ScaleMP has developed a software technology named vSMP, which is described and claimed in the Application. This

technology provides a shared virtual machine across multiple independent physical machines. vSMP aggregates resources of the underlying physical hardware, providing a single process space, memory address space, and I/O space for an operating system. As a result, multiple physical machines operate as a single virtual machine with the aggregated number of CPUs, memory size, and I/O space. This is a different, novel, and unique form of virtualization of resources, relative to methods that were previously known in the art.

6) Virtualization techniques prior to the development of vSMP subdivided a single computer into multiple smaller computers, providing instruction set emulation if required, as well as I/O virtualization. Such techniques replace direct access to the underlying hardware with a simplified virtual device that a "guest" operating system can use. This form of virtualization, as described in U.S. Patent 6,075,938 and in the "Disco" article cited against the Application (both by Bugnion), thus subdivides a single machine into many smaller "virtual machines," which have some number of virtual processors, some amount of memory, and some access to I/O and network resources. This is precisely the type of virtualization currently used in VMWare and similar products, and it is diametrically opposite in functionality and intention to vSMP.

7) The methods of virtualization described by Bugnion and implemented in VMWare do not allow the individual virtual machines to share memory, in the sense that for any two of these virtual machines, we cannot run a single process that consumes the sum of all memory across these machines. They also do not allow the

individual virtual machines to share processors, in the sense that the operating system on running on one virtual machine cannot directly schedule work on the second virtual machine. vSMP provides these capabilities, and does so transparently, permitting a standard operating system, without change, to run over multiple separate and independent computers.

8) The two different forms of virtualization I have described above – Bugnion's virtualization (VMWare) as opposed to that described in the Application (vSMP) - are orthogonal to each other. By "orthogonal," I mean that it is not an obvious or natural step to obtain the one form from the other form. Specifically, while it may be natural to use certain techniques and methods to <u>subdivide</u> one machine into many machines, it is unnatural to use the techniques of subdivision to <u>aggregate</u> multiple machines. While both forms use the term "virtual machine," the term is being used in a different technological context in each case. A "virtual machine" is not necessarily a physical machine to be sure, but it has a very different meaning depending upon which virtualization method is in use.

9) Moreover, the use of the term "virtual machine" in US Patent Application Publication 2003/0005068 (Nickel) is orthogonal to both of these concepts. Nickel relates to the venerable Parallel Virtual Machine (PVM) system developed decades ago, which has nothing to do with either of the virtualization concepts discussed above. The term "Virtual Machine" is, in the PVM context, a descriptive name, not a set of techniques to create multiple machine instances out of a single machine instance (as in VMWare), nor a set of techniques to create a seamless single

system image out of many machines, enabling all machines to transparently use all resources of each machine (as in vSMP). PVM is simply a programming framework for application-level programming, which allows single or multiple processes on each independent machine to communicate in a predefined manner. This framework allows the programmer to create a distributed algorithm for computing, using a number of networked machines, each running its own processes.

10) In the context of PVM programming, the term "virtual machine" refers to the process or processes that the program creates using the independent operating system on each machine, enabling the programmer to send data back and forth between the machines in a well defined manner. A PVM of the type described by Nickel is incapable of running an operating system and can run only programs that have been suitably modified. Despite using the term "virtual machine," Nickel's form of "virtualization" is entirely different from both VMWare and vSMP.

11) The following table summarizes the differences between the three types of "virtualization" that I described above:

Feature	**VMWare**	**vSMP**	**PVM**
Transparently use all processors on multiple component machines	No	Yes	No
Transparently use all memory on multiple component machines	No	Yes	No
Transparently use all networking resources on multiple component machines	No	Yes	No

Make one component machine appear as more than one machine	Yes	No	No
Enable unaltered programs to run utilizing resources	Yes (in most cases)	Yes	No
Enable unaltered operating systems to run utilizing resources	Yes (in most cases)	Yes	No
Enable large machine count single system images	No	Yes	No
Enable distributed programs to run in processes	Yes	Yes	Yes (must be written to PVM API)

It can be seen in the table above that aside from their use of the term "virtual machine," the three technologies have little in common. PVM seeks only to solve the bare-minimum issues of how to launch a collection of N distributed processes and how to provide a communication layer between them. Unlike vSMP, VMWare seeks to subdivide resources, not aggregate them, and to hide full system resources from operating systems, rather than exposing them.

12) For these reasons, a programmer in the field of multiprocessor computing - the "person having ordinary skill in the art" with respect to the Application - would never have even thought to apply the principles of process communication used in PVM in order to modify a VMWare virtual machine monitor. PVM and VMWare address different types of problems in ways that are mutually orthogonal, as I have explained above. Collaboration of multiple, loosely-coupled physical machines by PVM runs diametrically against the VMWare principle of subdividing a single physical machine into multiple virtual machines.

13) Furthermore, even if such a programmer had been "inspired" by PVM to try to modify the VMWare software in an attempt to run a single virtual machine over multiple physical machines, he still would have had no idea how to do so. As the table above shows, vSMP has a range of features that are outside the scope of capabilities of either VMWare or PVM. Solving the problems related to implementing these features required the inventors of vSMP to make a number of non-obvious inventive steps, as explained and claimed in the Application. The prior art does not provide the teachings that would have been needed by the person of ordinary skill in order to make these steps.

14) In the context of multi-computer systems, vSMP solves a specific set of problems that the other forms of virtualization do not address (and in some cases even exacerbate). By aggregating resources to provide a single system image over across multiple physical machines, vSMP provides a single point of management, a single operating system, and a single I/O space. This allows for:

- A reduction in system management costs and complexity, with only one "machine" to manage.
- A reduction in cost of operating system licenses: only one license required, rather than many.
- A simplified operating environment, enabling programs to operate in parallel without any additional message-passing layer (as in PVM).
- The ability to seamlessly use all system resources without modifying the operating system or application code.

15) In this respect, vSMP addresses a need that has long been felt in the field of high-performance computing (HPC): How to achieve the computing performance level of a supercomputer without high-cost dedicated hardware and special-purpose software? vSMP solves this problem by enabling multiple, generic, low-cost computers to be coupled together efficiently to create a single (virtual) high-power machine, in the manner that is described and claimed in the Application. Until very recently, VMWare virtualization was not even used in HPC, and its use in this area is still very limited due to the overhead of the virtualization process. vSMP has overcome these limitations and is rapidly growing in use and recognition.

16) The unique advantages of the vSMP solution have led to its adoption in HPC products offered by a number of industry leaders. For example, Exhibits B and C, attached hereto, contain product literature describing vSMP-based products that are distributed by Hewlett-Packard and Dell. Other major manufacturers offering vSMP-based products include IBM, Sun Microsystems, Cray, and Silicon Graphics. Exhibit D contains a press release in this regard by Cray.

17) The success of vSMP has led to the emergence of at least one imitator: 3Leaf Systems (Santa Clara, California). The 3Leaf product is described in a white paper entitled, "Next Generation Data Center Environment for HPC," which is attached hereto as Exhibit E.

18) Thus, to conclude, vSMP, as described and claimed in the Application, satisfies a long-felt need in the HPC industry and has enjoyed substantial market acceptance as a result. The success of vSMP has begun to attract imitators to the market. On the basis of these objective indicators, it is clear that the invention claimed in the Application is non-obvious. These secondary considerations are in addition to my analysis above, in which I showed that a person having ordinary skill in the art would anyway have been unable to derive the claimed invention from the Bugnion and Nickel references.

19) I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and conjecture are thought to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application of any patent issued thereon.

Dr. Joseph Landman, Citizen of USA
2433 Woodmont East, Canton, MI 48188

Date: 13-October-2009

EXHIBIT
tabbies
A

Joseph I. Landman,
2433 Woodmont East,
Canton, MI 48188
landman@scalableinformatics.com

EDUCATION:

Wayne State University, Detroit, Michigan
Department of Physics and Astronomy,
Ph.D. in Physics (Computational Theoretical), December 1997
Dissertation: *A Molecular Dynamics Investigation of Low Temperature Grown GaAs*
Advisor: Dr. Caroline Morgan

Michigan State University, East Lansing, Michigan
Department of Physics and Astronomy,
M.S. in Physics, June 1990

SUNY Stony Brook, Stony Brook, New York
Department of Physics,
B.S. in Physics, Dec 1987

HONORS AND AWARDS:

SGI Excellence in Systems Engineering, 1997
New York Regents College Scholarship 1983

RESEARCH AND PROFESSIONAL EXPERIENCE:

Cofounder of Diagnaid Inc Jan 2009-present
Building business model and plans, investor materials. Business and technology planning, marketing planning, sales planning and projection. Pursuit of funding from private capital markets, SBIR. Private capital market pitches and discussions, due diligence processes.

Cofounder of Micass LLC/Navivus Inc Jan 2005-Jan 2009
Building business model and plans, investor materials. Business and technology planning, marketing planning, sales planning and projection. Pursuit of funding from private capital markets, SBIR, MTTC. Prototyping efforts, proofs of concept. Market research on life science computing, medical image processing and related markets. Marketing events, one-on-many presentations and discussions. Private capital market pitches and discussions, due diligence processes.

Research Assistant Professor, Wayne State Jan 2005-June 2007
Teaching high performance computing programming methods to graduate students. Working with a research group on medical image processing tools, and applications around volumetric data navigation. Invention disclosure and due diligence on a novel and intuitive navigation method for clinicians. Began patent application process.

Founder and CEO of Scalable Informatics Inc. Aug 2002-present
Designing and building scalable storage and computing systems. Business development and marketing activities associated with growing a business. Partnership management, technical support, and engineering/research efforts.

Senior Scientist, Bioinformatics, MSC Software Apr 2001-Aug 2002
Architected a distributed parallel computing environment for Bioinformatics applications (MSC.LIFE^{TM}). Built a technical group responsible for the implementation of this product. Educated customers on the product. Performed benchmarking functions to demonstrate scalability. Designed and implemented cluster hardware for internal research and development efforts as well as for customers. Problem resolution efforts, documentation, and analysis. Collaboration with various customers on research projects of mutual interest, including distributed data mining, bioinformatics visualization and information representation.

Systems Engineering Specialist, SGI Apr 1995-Apr 2001
Architected and implemented numerous applications for internal and customer use, including some SGI product (SGI GenomeCluster^{TM}, a scalable bioinformatics computing resource) and pre-cursors (Roboinst derived from my Autoinst). Built collaborations with research universities and individual research groups. Parallelized codes for shared and distributed memory machines. Benchmarked large scale computational systems. Researched performance issues with application scaling for a variety of architectures. Authored papers with collaborators. Collaborated on educational initiatives including Wayne State University's IGERT for Computational Science, and the associated institute ISC. Gave talks at scientific meetings and trade group meetings. Consulted on high performance computing issues for SGI research and educational customer base. Helped customers with algorithm re-design for high performance, re-implementation, and general assistance with coding issues. Performed large scale visualization of proteomics data. Performed proof-of-concept research for customers, including data mining on 10^9 record databases.

Consultant Jan 1992-Apr 1995
IT and infrastructure consultant for a variety of customers.

Graduate Research and Teaching Assistant Sept 1990-Apr 1995
Investigated physical and chemical models using simulation techniques on

supercomputers. Built programs to visualize molecular dynamics trajectories, analyse large runs, recover from problems within the runs. Designed algorithms to compute volumetric charge distribution from modified self-consistent *ab-initio* molecular dynamics code. Parallelized code for SMP SGI machines, vectorized code for Cray machines. Developed new programs and anlyses by encapsulating existing molecular dynamics codes within method/procedure calls to provide a Murnaghan equation of state data fit. Built a code to generate an STM image from a slice through the supercell volume. Managed students, computers, and information access/security. Taught lecture courses, as well as lab and recitation sections of lecture courses.

Graduate Research and Teaching Assistant Sept 1988-Jun 1990
Investigated physical models using simulation techniques on mini-computers. Built programs to visualize physical phenomenon, and measurements. Taught lecture courses, as well as lab and recitation sections of lecture courses.

Associate Engineer, IBM T.J. Watson Research Center Jan 1988-Aug 1988
Performed measurements of onset temperature and critical current density for high Tc superconductors. Built programs to automate measurements using IEEE 488/GPIB bus. Built programs to automate data analysis, reporting, and identify anomolous results for further study. Removed data collection bottleneck resulting in the experiment frequency being dominated by the cryogenic system cooling rates.

Undergraduate Research Assistant Jun 1984-Dec 1987
Worked on developing a computer control for the laser cooling of atomic beams experiment at SUNY and the National Institute of Science and Technology (The PI of this effort at the NIST won the 1997 Nobel prize in Physics for this research, see http://www.nist.gov/public_affairs/releases/n97-26.htm).

TEACHING EXPERIENCE:

Scalable Informatics: sales and marketing training, courses on high peformance computing programming

MSC Software: sales and marketing training. End user training. HPC tutorials.

SGI: sales and marketing training. End user training. HPC tutorials. Data Mining for bioinformatics research.

Wayne State: Introductory (calculus based) Physics lecture, recitations, and laboratories. Advanced laboratory sections in modern physics experiments.

Recitation sections for non-calculus based physics.

Michigan State University: Advanced laboratory section for modern physics experimental course.

PATENTS:

- USPTO #7,249,357 "Transparent distribution and execution of data in a multiprocessor environment "

PUBLICATIONS:

- **"Accelerating HMMer searches on Opteron processors with minimally invasive recoding",** Joseph I. Landman, Joydeep Ray, J. P. Walters: AINA (2) 2006: 628-636
- **"Parallelization of a Legacy Program using OpenMP",** Landman J. and Piecuch P., ACM FORTRAN Forum, 19, 2, August 2000, 16-23.
- "**Parallelization of a multi-reference coupled-cluster method**" Piotr Piecuch and Joseph I. Landman, *Parallel Computing*, Vol 26, 7-8, July 2000
- **"Arsenic Interstial Pairs in GaAs"** P. Papoulis, C.G. Morgan, J.T. Schick, J.I. Landman, and N. Rahhal-Orabi, Materials Science Forum **258-263**, 923 (1997).
- **"Arsenic Interstials and Interstitial Complexes in Low-Temperature-Grown GaAs"** J.I. Landman, C.G. Morgan, J.T. Schick, P. Papoulis, and A. Kumar, Phys. Rev. B **55**, 15581 (1997).
- **"Arsenic-Antisite-Related Defects in GaAs Grown at Low Temperatures: Characterization of Localized States"** J.I. Landman, C.G. Morgan, J.T. Schick, A. Kumar, P. Papoulis, and M.F. Kramer, Materials Science Forum **196-201**, 249 (1995).
- **"Antisite-Related Defects in GaAs Grown at Low Temperatures"** J.I. Landman, C.G. Morgan, and J.T. Schick, Phys. Rev. Lett. **74**, 4007 (1995).
- **"Interstitial Defects in II-VI Semiconductors: Role of the Cation d States"** J.T. Schick, C.G. Morgan, and J.I. Landman, Materials Science Forum **83-87**, 1253 (1992)

PRESENTATIONS:

- "**The unreasonable effectiveness of clusters for life science and informatics computing**", ClusterWorld 2004 invited talk
- "**Building Software for High Performance Informatics and Chemistry**", ClusterWorld 2003 invited talk

- **"Bio-IT: the nuts and bolts",** Eastern Michigan University Center for Entrepreneurship and Michigan Center for Biological Information, Jan. 2003

- **"Scalability Matters - Why We Need to Make Bioinformatics Programs Scalable, and Results from Work on Various Programs",** Michigan State University, Center for Biological Modeling, Oct. 2001

- **"The DNA of Computing"** Intel invited talk to LinuxWorld August 2001, on High Performance Informatics,

- **Building Optimal Computational Environments for Bioinformatics and Computational Chemistry",** SGI Bioinformatics Seminar, ISMB 2000, August 2000

- **"Linux for Software Development**", SGI Linux University, March – July 2000

- **"Linux Clusters for Computational Science**", Semana de Supercómputo, UNAM Departamento de Supercómputo , May 2000

BOARD MEMBERSHIP:

Honorary (Emeritus) Director: Bioinformatics.org

EXHIBIT

tabbies®

B

HP and ScaleMP put you in charge

All-in-one "Shorty" with ScaleMP vSMP Foundation

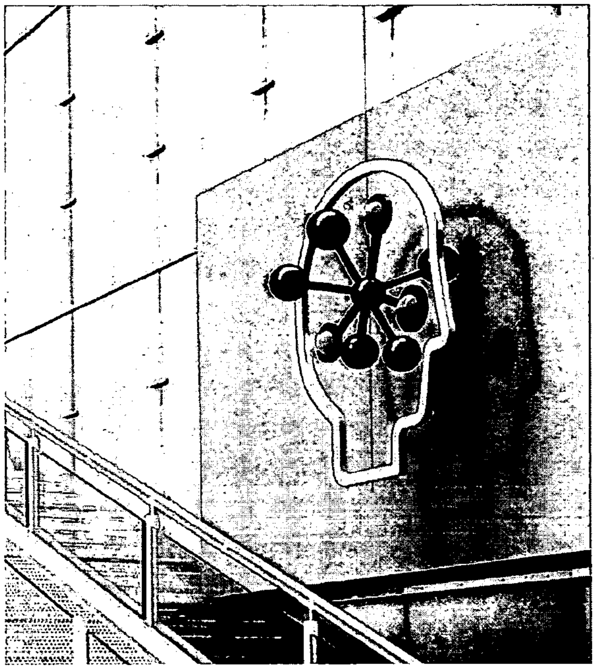
A supercomputer-in-a-box workgroup solution makes HPC simple and personal.

Affordable HPC at your command
A High Performance Computing (HPC) capability is essential for organizations to compete in a demanding global economy. Clustering has helped make HPC more affordable and accessible for parallel workloads, but remains complex. In addition, solutions for workloads requiring large memory have been too expensive until now.

HPC data center capability, outside the data center
HP and ScaleMP extend the affordability and accessibility of clusters to large memory workloads, while simplifying administration. ScaleMP vSMP Foundation seamlessly aggregates multiple industry-standard x86 systems into a single virtual system. The combination of vSMP Foundation and the HP Cluster Platform Workgroup system ("Shorty") enables workgroups and small organizations to deploy high performance capability outside the data center, running mixed large memory and parallel workloads. The "Shorty" utilizes HP BladeSystems in a small, compact design that conserves space while decreasing power consumption and heat generation.

Solution benefits

- Simple and rapid deployment on a compact, powerful system via Cluster Platform Express
- Run parallel applications on up to 128 cores
- Run extremely large jobs and models with up to 1 TB of shared memory
- Single system simplicity
- Cost-effective, power-efficient and flexible HP BladeSystem design



Working together with HP
HPC solutions enable rapid advancements innovation, cost-efficiency and productivity. Using proven HPC methodologies, organizations of all sizes can speed and optimize their engineering, financial, drug discovery and development processes to drive greater success. Working collaboratively with our partners, we apply our vast knowledge and experience in HPC to deliver complete solutions that help you tap into the strength and flexibility of supercomputing powered by highly reliable HP servers.

ScaleMP™ | hp®

Single virtual HPC system

vSMP Foundation, deployed on the HP Cluster Platform Workgroup system, offers tremendous price/performance advantages for the HPC market. In essence, it provides a unique way to leverage entry-level systems to reduce the total cost of ownership (TCO). It delivers the operational simplicity of traditional shared-memory systems while keeping the acquisition cost associated with clusters. With an ability to support both parallel and large memory workloads, the solution is a supercomputer-in-a-box. The HP Cluster Platform Workgroup system, leveraging the top selling HP BladeSystem technology, requires no special power, cooling or staff and can be deployed outside the data center. With a footprint of less than two square feet, it can deliver almost a TFLOP of power, with up to eight nodes using HP ProLiant BL460c Servers, all communicating over the integrated InfiniBand network incorporated into the workgroup system.

HP products supported
Our solutions are built on best-selling HP ProLiant servers, world-renowned for reliability and availability. From server blades to clustered servers, HP ProLiant servers provide the utmost confidence for your business. With new Intel® Xeon™ based HP ProLiant server models, HP further extends the advantages of x86 computing—delivering more cost-effective, industry-standard solutions for applications requiring expanded memory and outstanding price/performance. HP ProLiant servers are your best choice for long-term dependability, flexibility and growth.

Operating systems supported
Our Linux-based solutions are built on innovative software and standards-based servers—delivered by service professionals with extensive experience. By working with HP, we can leverage its worldwide leadership in Linux to provide optimized solutions based on HP Open Source Integrated Portfolio (OSIP) and HP Open Source Middleware Stacks (OSMS). You can trust our combined expertise to provide the proven, cost-effective Linux solutions you need to drive a fast return on your IT investments.

Service and support
ScaleMP solutions build on HP BladeSystem, the foundation for HP Cluster Platform Workgroup Systems, preconfigured, turn-key HPC clusters available from HP and our partners. ScaleMP installation services and support are available directly from ScaleMP and its partners.

Building on the value of strong relationships
By working collaboratively with HP, you can leverage extensive resources, deep experience and broad industry knowledge to provide innovative solutions that drive positive business results and long-term value. With a proven record of success in virtually every industry in every region worldwide, HP understands what you need to increase your organization's success. Together, we can deliver best-fit technology and services to meet your unique business requirements.

ScaleMP is a leader in virtualization for high-end computing, providing higher performance and lower TCO. Using software to replace expensive and proprietary symmetric multiprocessing systems (SMP), ScaleMP offers a new computing paradigm, aggregating industry-standard systems into a single virtual x86 system to tackle the most difficult workloads.

Working together to drive success
By combining standards-based and high-performance technologies from HP with our leading-edge applications and comprehensive services, we can deliver a powerful, flexible solution that meets your technology requirements. With a tailor-made solution designed by business partners you trust, you can face your toughest business challenges—including increasing productivity, quickening time to market and reducing operating costs—through greater efficiency, rapid discovery and reduced development cycles.

Technology for better business outcomes

To learn more, visit www.hp.com
www.scalemp.com

4AA2-3301ENW, November 2008

ScaleMP™ | hp®

EXHIBIT

C

SIMPLY PERFORM!

TIRED OF WAITING FOR SHARED DATACENTER RESOURCES?

TIRED OF WAITING FOR YOUR SIMULATIONS TO COMPLETE?

WOULD YOU PREFER PROGRAMMING WITH OPENMP?

WANT TO INSTALL A SUPERCOMPUTER IN LESS THAN A DAY?

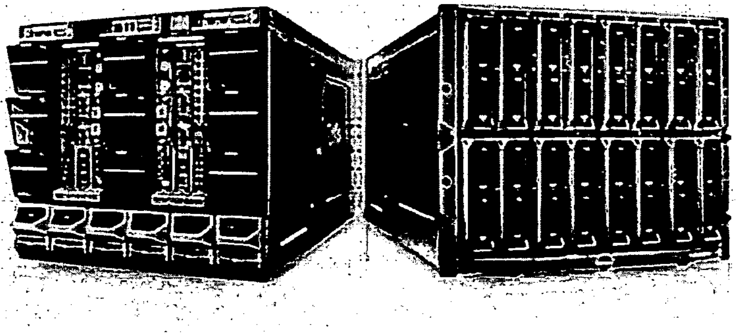
DREAMING ABOUT INFINIBAND SPEED WITHOUT HEADACHES?

WANT TO SIMPLIFY CLUSTER STORAGE BY A FACTOR OF 16?

Dell PowerEdge M1000e

Single Virtual System

Optimized for High Performance Technical Computing (HPTC)



1.6 TFLOPS
+ 128 Cores
+ 3TB Shared Memory
+ 19TB Internal Storage

1 Operating System

DELL

Dell PowerEdge M1000e Single Virtual System At A Glance

The Dell PowerEdge M1000e Single Virtual System is an x86 supercomputer with up to 32 processors (128 cores) and up to 3TB of shared memory running a single instance of a Linux operating system. Based on ScaleMP's vSMP Foundation™ software that creates a single virtual system by aggregating multiple Dell MxxO blade servers within a M1000e chassis, this system is ideally suited for high performance computing applications in the financial services, life sciences, engineering and educational institutions. It offers significant price/performance advantage, power consumption savings, and higher density over traditional and proprietary SMP systems. Finally, high performance SMP's are affordable again...

PROBLEMS SOLVED

Reducing cost of traditional SMP Systems

Traditional high-end x86 systems with four, eight or more sockets are based on lengthy and proprietary R&D developments that must be passed on to end users. These systems also usually incorporate older generation components and chip speeds. This results in expensive solutions that have lower compute density, consume more power, and cost more on a per-socket basis compared to dual-socket systems. ScaleMP's vSMP Foundation, by aggregating the more powerful dual-socket servers into a single virtual system offers better performance, scalability, yet at a lower cost!

Reducing complexity of cluster deployments

Today's clusters are designed to provide high-density coupled with excellent performance and power efficiency. However, management costs are high: multiple Operating Systems required, replication of applications and content. In addition, a complex high-speed cluster file-systems or proprietary external storage solutions must be implemented. Applications are limited to the memory footprint per system. ScaleMP's vSMP Foundation converts clusters into affordable SMPs: single Operating System, internal storage and large memory. Same components – just simpler to manage and run.

BENEFITS

Large memory

The Dell PowerEdge M1000e Single Virtual System enables an application to use the aggregated memory of all the blades in the system. In the extreme, a single application process can leverage up to 3TB RAM. Large memory also reduces the need to use external high-performance storage systems for swap or scratch space. Application runtime is dramatically reduced by running simulations with in-core-solvers or by using memory instead of swap for large models. In addition, both traditional SMP codes (OpenMP) and distributed applications (MPI) run at optimal performance on the same physical infrastructure.

Compute & memory intensive applications

For workloads that require a high core count coupled with shared memory, users have traditionally acquired proprietary shared-memory systems. The PowerEdge M1000e Single Virtual System provides a very cost effective x86 alternative to these expensive RISC systems. As opposed to traditional SMP or NUMA architecture where memory bandwidth decreases as the machine scales, it combines memory-bandwidth across boards and demonstrates close to linear memory bandwidth scaling. The scalable memory bandwidth delivers excellent performance for memory intensive applications.

Ease of use

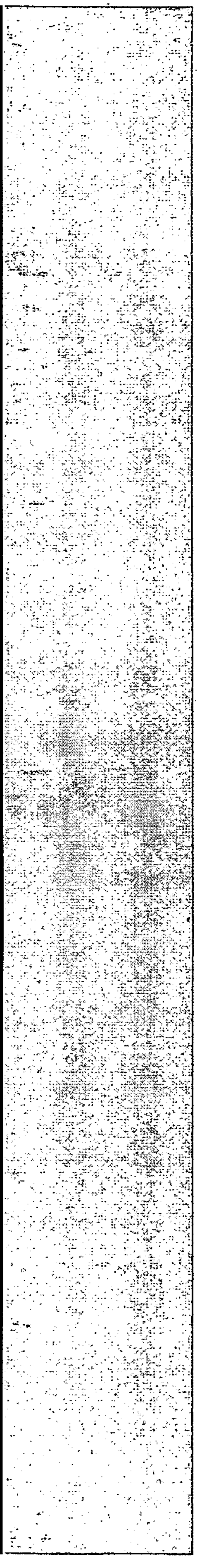
For workloads that otherwise require a scale-out approach, the Dell PowerEdge M1000e Single Virtual System provides ease of use by having a single system to manage, compared the complexities involved with managing a cluster. A single system removes the need for cluster file systems, cluster interconnect issues, application provisioning and installation and update of multiple operating systems and applications. This results in significant savings at installation, and ongoing operations.

Simplified I/O architecture

High-bandwidth I/O requirements in a scale-out model can be complex and costly, usually involving HBA's, and FC switch infrastructure. The Dell PowerEdge M1000e Single Virtual System aggregates each individual server's network and storage interfaces. I/O resource consolidation reduces the number of drivers, HBA's, NIC's, cables, and switch ports and all the associated maintenance overhead. The user needs fewer I/O devices to purchase, manage and service.

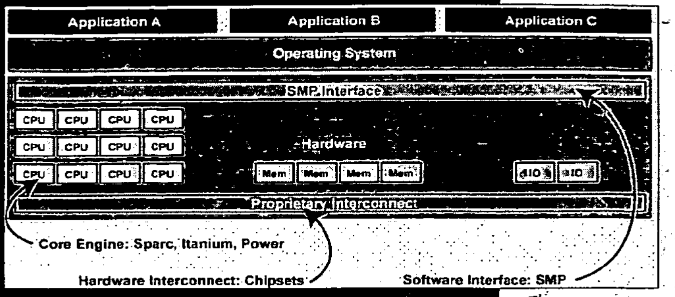
Improved Utilization

For large compute farms deployments the Dell PowerEdge M1000e Single Virtual System becomes an attractive alternative for organizations that need to run hundreds or thousands of simulations at once. As opposed to hundreds of servers, where each server operates at 80 percent utilization (to allow for runtime peaks), fewer larger systems can run more applications on the same footprint.



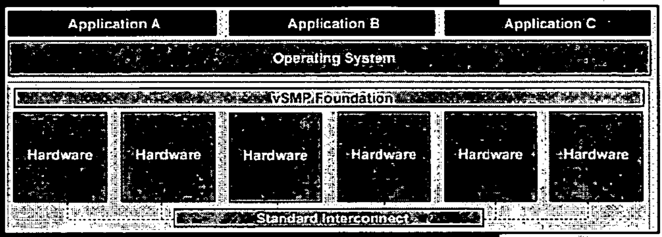
How Does It Work?

Traditional SMP systems run a single operating system (OS) which interacts with the system using a well-defined hardware interface. This interface provides the OS with predefined services to use and control the hardware, including hardware detection and probing, memory ordering semantics, I/O space access and interrupt delivery mechanisms. An example of such hardware interface is the Intel's Multi-Processor Specification (MP Spec) which defines a standard interface between the hardware and the OS to make it easy for the OSVs and OEMs to quickly support a wide range of platforms with one OS version, a benefit they already enjoy in the Uni-processor desktop market for Intel Architecture CPUs. In essence, the MP Spec brings the same "shrinkwrap" benefits of the desktop market to the MP market. For a traditional SMP system, such interface is implemented in a silicon chipset. In addition to the hardware interface, an SMP system consists of CPUs, memory and I/O subsystems, all connected together with a proprietary backplane or interconnect such as Intel's FSB (Front Side Bus), AMD's HT (Hyper-Transport), SUN's CrossBar SGI's NUMALINK and IBM's XA. The proprietary backplane (system interconnect) is where today's SMP systems differ the most from one another.



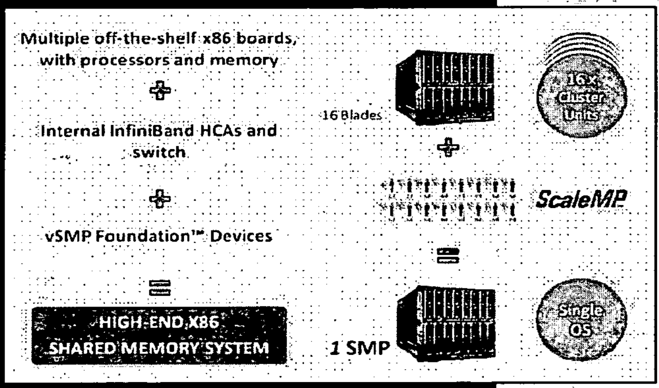
THE SCALEMP VERSATILE SMP™ (VSMP) ARCHITECTURE

The vSMP architecture utilizes off-the-shelf components and does not require any custom parts. Its key value is the utilization of software to provide the chipset services that are otherwise required in creating traditional multi-processor systems. vSMP Foundation provides cache coherency, shared I/O and the system interfaces (BIOS, ACPI) , which are required by the OS. *The vSMP architecture is implemented in a completely transparent manner*; no additional device drivers are required and no modifications to the OS or the applications are necessary.



From a hardware perspective, vSMP Foundation requires:

- Multiple Dell x86 systems or blades
- InfiniBand HCA's, cables and switch to interconnect the systems or blades
- vSMP Foundation Devices – Flash-based storage devices (one per board/system) with the appropriate vSMP Foundation product supporting the specific Dell products


Multiple off-the-shelf x86 boards, with processors and memory
+
Internal InfiniBand HCAs and switch
+
vSMP Foundation™ Devices
=
HIGH-END X86 SHARED MEMORY SYSTEM
16 Blades
16 x Cluster Units
+
ScaleMP
=
1 SMP
Single OS

From a system perspective, vSMP Foundation provides:

- One single system: once loaded in memory of each system boards, vSMP Foundation aggregates all the resources of the multiple physical systems, initializes the interconnect fabric, and creates the required BIOS and ACPI environment to provide the OS a coherent image of a single virtual system. vSMP Foundation then uses a software-interception engine in the form of a Virtual Machine Monitor (VMM) to provide a uniform execution environment.
- Coherent Memory: vSMP Foundation maintains cache coherency between the individual boards using multiple advanced coherency algorithms that operate concurrently on a per-block basis, based on real-time memory activity access patterns. vSMP Foundation leverages board local-memory together with best-of-breed caching algorithms to offset the effect of interconnect latencies.
- Shared I/O: vSMP Foundation aggregates I/O resources across all boards into a unified PCI hierarchy and presents them as a common pool of I/O resources to the OS and applications

Frequently Asked Questions

Question	Answer
What node is the master node?	There is no concept of a master node in shared memory systems (as there is only sa ingle node). vSMP Foundation however has a concept of primary device (inserted in the first board which contains the configuration information, and is where the system keyboard, video and mouse should be used) and secondary devices (for all other boards).
Where does the operating system run?	The operating system runs over the entire system. It can boot from local drives or from the network.
Do applications or operating systems need to be changed or modified to run on the system?	The systems run standard Linux operating systems distributions. Any x86 binary runs as-is, exactly as if it were running on the standard dual socket x86 server. However, as vSMP Foundation enables systems with up to 128 processor cores and terabytes of RAM, several tuning and optimization should be considered to reap maximum performance. ScaleMP offers application execution guidelines for significant number of applications as well as tuning suggestion for the Linux kernel.
Can I disable vSMP Foundation to run MPI?	vSMP Foundation can be certainly be disabled if one wants to run native cluster. However, as MPI applications run at the same performance level with, or without vSMP Foundation, most users just run MPI applications in shared memory.
How stable is this technology?	vSMP Foundation first appeared on the market in 2005, and is now implemented in production at over 100 sites worldwide. End-users include manufacturing companies, higher education institutions, life-sciences and pharmaceutical organizations as well as leading financial institutions.
What is the latency for off-board memory access?	The ScaleMP versatile SMP (vSMP) architecture is hybrid COMA-NUMA architecture. As such there is no notion of fixed latency for memory access. In essence, the vSMP architecture minimizes the number of times a processor fetches memory from a remote board. Think of it as additional, board-level, cache.

Benefits By Industry Sectors

Manufacturing

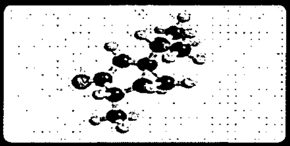
Leverage high memory bandwidth and large number of cores for structural and impact analysis, and computational fluid dynamics applications. Take advantage of large shared memory for implicit analysis, pre-processing and post-processing.

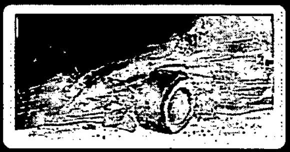
Energy

Most reservoir and volume interpretation applications require large memory and high memory bandwidth. Seismic processing requires a large number of processors. The vSMP architecture is ideal for such applications.

Numerical simulations

Flexibility to run numerical simulations using all the memory in the system, multiple processes in parallel sharing the memory, or one or more jobs running in multi-processors mode.





Higher education and research

Dynamic adjustment to the mix of multidisciplinary applications, as well as ever-changing research priorities; from jobs that require a large memory footprint, high number of processors, or small to medium simulations in throughput mode.

Life sciences

Flexible, high-performance system to run a large number of disparate legacy, OpenMP, MPI applications in one system, by leveraging high number of processors, large memory or bandwidth or a combination thereof.

Electronic Design Automation (EDA)

Utilize the same hardware infrastructure for large shared memory processing during validation phases (prior to tape-out) and running multiple concurrent user jobs on large core count in day-to-day use.

Bottom Line: Get A Headstart On The Competition

End user	Rubber Manufacturing Corp	Engineering Services Company	Formula 1 Team
Existing infrastructure	High-core count Itanium system	Multiple 2-socket workstations	Large Memory Itanium system
Challenges	Need to run thread-based applications, such as Gaussian, as well as MPI-based Computational Chemistry applications	• Existing models (Abaqus) grow fast and no longer fit engineers' workstations • Need to run large simulations in batch at night • No in-house skills to run x86 InfiniBand cluster and cannot afford RISC systems	• Need to generate large meshes as part of pre-processing of whole car simulation (FLUENT TGrid) • Mesh requirements are over 200 GB in size • Wants to standardize on x86 architecture due to lower cost
Solution	• 8 Xeon Quad-core processors • 32-cores • 128 GB RAM	• 8 Xeon Dual-core processors • 16 cores • 128GB RAM	• 24 Xeon Processors • 96 cores • 384 GB RAM
Performance Benefits	• Significantly faster than the existing IA64 SMP • Performance is comparable to cluster performance with similar-hardware	• Significantly faster than existing workstations • Performance is comparable to cluster performance	Evaluated and proven to be faster than alternative systems (x86 and non-x86)
Operational Benefits	No IT resources required for day-to-day operation	No IT required for day-to-day operation	Similar to existing system
Capital Expenditure Benefits	Significant savings compared to upgrade or replacement of existing system with IA64		Significant savings compared to existing and alternative systems considered
Versatility		Interactive jobs during the day, batch jobs at night	Used for large memory jobs (TGrid) and regular FLUENT (MPI) solvers on large number of cores
Investment protection		Expected to double the resources	Headroom for growth

Available Configurations

Specifications

- Min. / Max. boards: 2 / 16
- Min. / Max. memory (GB) per board: 4 / 256
- Min. / Max. processors per board: 1 / 4
- Min. / Max. cores per board: 1 / 16
- Max. system memory: 4 TB
- Max. system processors: 64
- Max. system cores: 128

Supported platforms: http://www.ScaleMP.com/spec

Screaming Performance

Facts

- The fastest x86 system by memory bandwidth
- The 15th fastest system by memory bandwidth globally
- The fastest x86 system by SPEC CPU2006
- The largest shared-memory x86 system by RAM size and core count

For More Information

Dell

- EMEA: Paul Brook (Paul_Brook@Dell.com) EMEA HPC Programme Manager
- USA: Karl Cain (Karl_Cain@Dell.com) HPC Business Development Manager

ScaleMP

- dell@scalemp.com

EXHIBIT
tabbies
D

THE SUPERCOMPUTER COMPANY

Print Page Close Window

News Release

Cray and ScaleMP Announce Strategic Alliance

SEATTLE, WA and CUPERTINO, CA, Mar 05, 2009 (MARKET WIRE via COMTEX) -- Global supercomputer leader Cray Inc. (NASDAQ: CRAY) and ScaleMP, a leading provider of virtualization solutions for high-end computing, today announced a strategic alliance to offer joint solutions based on the Cray CX1(TM) deskside supercomputer and ScaleMP's vSMP Foundation. Available immediately, the joint solution will target the High Performance Computing (HPC) segment allowing customers to operate a shared-memory, deskside supercomputer that scales up to 128 cores and 1TB of shared memory.

The Cray and ScaleMP strategic alliance is focused on enabling supercomputing at the workstation level. The combined Cray CX1 system and vSMP Foundation solution enables workgroups and small organizations to deploy high performance computing capabilities that harness the power of multiple processors while simplifying their operational environment. This solution is versatile, able to run a variety of Linux(R) workloads such as large memory, parallel workloads and high core count shared memory applications, and delivers excellent performance across many programming models ranging from MPI, OpenMP and legacy code.

"Cray and ScaleMP are addressing important requirements for HPC by offering a personal supercomputer workstation," said Earl Joseph, IDC Program Vice President. "Many departmental and work group users of HPC applications have been constrained by the lack of in-house skills to move up to clusters from workstations. This solution will allow these users to more easily scale up their simulations and models and boost productivity and competitiveness without the added complexity."

"I am very excited about our collaboration with Cray," said Shai Fultheim, founder and CEO of ScaleMP. "Cray is synonymous with excellence in the high-end supercomputer segment. This announcement enables HPC customers to get Cray performance at the deskside, in a cost-effective, workstation-like simplicity. Cray customers will be leveraging the capabilities of vSMP Foundation to achieve a flexible compute resource capable of solving bigger problems -- accelerating time to market and innovation."

"The ScaleMP vSMP Foundation virtualization software is an excellent fit for the Cray CX1, which was designed specifically to harness HPC for individuals and departmental workgroups," said Ian Miller, senior vice president of the productivity solutions group and marketing at Cray. "By creating a single shared memory virtual system, the joint solution can now support large memory and large core count workloads in addition to parallel workloads, while simplifying the installation and management of the system."

vSMP Foundation aggregates multiple industry-standard, off-the-shelf x86 servers (rack mounted or blade systems) into one single virtual high-end system for the HPC market. vSMP Foundation provides customers with an alternative to traditional expensive symmetrical multiprocessor (SMP) systems and also offers simplified clustering infrastructure with a single operating system. It currently allows customers to create a single virtual SMP system with up to 32 sockets (128 cores) and up to 4 TB of shared memory in an energy-efficient, dense package.

About Cray Inc.

As a global leader in supercomputing, Cray provides highly advanced supercomputers and world-class services and support to government, industry and academia. Cray technology enables scientists and engineers to achieve remarkable breakthroughs by accelerating performance, improving efficiency and extending the capabilities of their most demanding applications. Cray's Adaptive Supercomputing vision will result in innovative next-generation products that integrate diverse processing technologies into a unified architecture, allowing customers to surpass today's limitations and meeting the market's continued demand for realized performance. Go to www.cray.com for more information.

Cray CX1 Supercomputer

The Cray CX1 product is an affordably-priced, deskside supercomputer. Easy to configure, deploy, administer and use, it is the "right size" in performance, functionality and cost for a wide range of users, from the single user who wants a personal supercomputer to a department of users as a shared clustered resource. Equipped with powerful Intel Xeon(R) processors and Windows(R) HPC Server 2008 or Red Hat Enterprise Linux with Clustercorp Rocks+, the Cray CX1 product offers performance leadership across a broad range of applications and standard benchmarks. For organizations wanting to harness HPC without the complexity of traditional clusters, the Cray CX1 supercomputer delivers the power of a high performance cluster with the ease-of-use and seamless integration of a workstation.

About ScaleMP

ScaleMP is the leader in virtualization for high-end computing, providing maximum performance and lower Total Cost of Ownership (TCO). The innovative Versatile SMP(TM) (vSMP) architecture aggregates multiple x86 systems into a single virtual x86 system, delivering an industry-standard, high-end symmetric multiprocessor (SMP) computer. Using software to replace custom hardware and components, ScaleMP offers a new, revolutionary computing paradigm. The company is backed by Sequoia Capital, Lightspeed Venture Partners, TL Ventures, and ABS Ventures. For more information, please call +1 (408) 342-0330 or visit www.ScaleMP.com.

Cray is a registered trademark, and Cray CX1 is a trademark of Cray Inc. All company and/or product names may be trade names, trademarks and/or registered trademarks of the respective owners with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

vSMP Foundation is a trademark or registered trademark of ScaleMP. All other brands or products are trademarks or registered trademarks of their respective holders and should be treated as such.

Cray Media:
Nick Davis
206/701-2123
nickd@cray.com

ScaleMP-Media:
Amar Rao
408/342-0330
PR@ScaleMp.com

SOURCE: Cray Inc.

mailto:nickd@cray.com mailto:PR@ScaleMp.com

tabbies EXHIBIT E

Next Generation Data Center Environment for HPC

Enabling the Dynamic Data Center

3LEAFSYSTEMS

Copyright

Notice: This document is for informational purpose only and does not set forth any warranty, express or implied, concerning any equipment, equipment feature, or service offered or to be offered by 3Leaf Systems. 3Leaf Systems reserves the right to make changes to this document at any time, without notice, and assumes no responsibility for its use. This informational document describes features that may not be currently available. Contact 3Leaf Systems for information on feature and product availability.

Export of technical data contained in this document may require an export license from the United States Government.

3Leaf Systems

3255-1 Scott Blvd., Suite 200

Santa Clara, CA 95054

Phone: 408.572.5900

Fax: 408.727.2008

www.3leafsystems.com

Virtualization and HPC

HPC solutions today face challenges in flexibility, cost, software development time and performance. In the enterprise world, some of these very same constraints have been solved by vendors with software virtualization solutions (e.g., VMware or Xen). However, the HPC market is different from the traditional enterprise market in that higher utilization of computing resources, a big problem for enterprise customers, is a distant second priority to performance. As such, virtualization solutions popular in the enterprise data center have little to offer to the HPC user, whose primary concern is performance or possibly performance per dollar or performance per watt.

Indeed, software virtualization solutions such as hypervisors attempt to multiplex multiple virtual machines onto a single physical machine to improve the physical machine's utilization. However, in the HPC environment, often a single job can fully utilize the machine. For many HPC applications, the problem is not machine utilization.

In addition, the focus on performance, performance per dollar, performance per watt, and flexibility has precipitated the trend towards cluster based computing. Cluster based computing takes advantage of inexpensive, commodity computing and network resources (e.g. x86 computers and high speed LANs) to form large pools of interconnected compute resources. Because a cluster can be sized to fit the application, these clusters have significant cost and flexibility advantages over the big iron supercomputers they replace.

Moving towards cluster based computing also means shifting to a new memory model. The memory model provided by many traditional super computers is coherent shared memory. The memory model presented by HPC clusters is distributed memory on independent nodes, connected by a low latency message passing network. The mapping of the traditional shared memory paradigm to that of distributed memory and message passing is not without difficulty. Some problems do not partition easily into computational threads that have low communication overhead. In others, while the partitioning may be obvious, the amount of communication or the latency presented by communication becomes a performance bottleneck.

Besides mapping the communication paradigm from shared memory to message passing, a programmer needs to deal with the hard resource constraints of the individual compute nodes when writing or porting an application to an HPC cluster. The compute nodes in traditional cluster computers have hard limits on the number of CPU cores available, the amount of memory and the amount of I/O.

3Leaf's Dynamic Data Center

3Leaf's Dynamic Data Center ("DDC") enables efficient scale-up computing across multiple commodity servers. This greatly expands the flexibility and scalability, enabling multiple physical servers to be dynamically grouped together into a single logical server. By allowing complete flexibility in how CPU and memory resources are provisioned, both capital and operational expenditures are significantly reduced.

3Leaf's unique technology for enabling CPU and memory virtualization includes both ASIC (the 3Leaf Coherent NIC) and software technology (the 3Leaf Distributed Machine Monitor). The 3Leaf Coherent NIC is the first of its kind, and extends the coherency domain of an x86 processor across multiple x86 commodity platforms using commodity switch fabrics. By doing this, the physical boundaries of a server are expanded, and it provides the backbone for the virtualization of compute and memory resources to efficiently span multiple x86 physical servers. In other words, for the first time ever, a single Guest OS can span across many servers to utilize CPU or memory resources to handle spikes in application traffic.

In order to fully address the x86 market with its CPU and memory virtualization technology, 3Leaf has licensed the processor interconnects from both AMD and Intel. The AMD processor interconnect is Coherent HyperTransport (cHT), and Intel's is QuickPath Interconnect (QPI). Licensing these core pieces of technology enables 3Leaf to support both AMD and Intel based solutions, and support 100% of the x86 commodity server market.

Specific to HPC, 3Leaf's DDC removes hardware barriers of existing HPC clusters, enabling:

- construction of coherent memory domains across an HPC cluster,
- partitioning of memory resources into multiple coherency domains, unrestricted by physical constraints (e.g. server blade), and
- partitioning of CPU resources across a cluster, allowing for the best allocation of CPU resources to fit the application.

3Leaf's ability to fully virtualize all components of an x86 server into pools that can span across multiple physical machines is truly game changing and sets 3Leaf apart from the competition. The Dynamic Data Center enables 3Leaf to provide a complete server virtualization solution for commodity platforms, allowing compute, memory, and I/O resources to be dynamically allocated and de-allocated as required.

3Leaf's Network Shared Memory

3Leaf's Dynamic Data Center also brings powerful shared memory concepts to the cluster computing environment. Nodes within a cluster can collectively create a coherent shared memory region (i.e., network shared memory). Every node within the cluster can contribute memory to the network shared memory area, which can be directly addressed by every other node that is part of the network shared memory cluster.

The result is that there can be up to one terabyte of shared memory between specific configurations ranging from 2 to 256 nodes in the cluster. This memory is directly addressable within the network cluster and allows all the operations that shared memory is capable of, enabling much more scalable and lower-overhead data sharing and communication methods than can be expected in a traditional cluster environment.

3Leaf network shared memory offers support for read-modify-write operations at the hardware level to shared memory addresses. This ability is very useful in a cluster environment where synchronization is much more costly, typically being carried out through expensive distributed locking methods.

3Leaf's Dynamic Data Center for HPC

3Leaf's view of virtualization offers new opportunities for cluster based computers to provide the advantages of traditional shared memory supercomputers while leveraging the cost and scalability advantages of commodity based cluster computing. 3Leaf's DDC allocates compute and memory resources as required by the job. Unlike traditional virtualization solutions, 3Leaf's solution does not multiplex physical resources amongst competing virtual machines.

The 3Leaf DDC allows cluster CPU resources to be allocated to virtual servers, without regard to the physical nodes on which the CPUs reside. Servers can be created that match the HPC application's requirements, instead of HPC applications having to conform to the physical machine's topology. In other words, the machine conforms to the problem, the problem does not have to conform to the machine.

By providing coherent shared memory across the nodes in the cluster, the 3Leaf solution offers the performance of traditional shared memory with the capabilities and configuration flexibility of a cluster based computer. In addition, applications that use message passing can benefit from a shared memory implementation of the message passing library (e.g. OpenMPI with shared memory messaging) without code changes. Indeed, a shared memory implementation of message passing need not incur the network stack and scheduling overhead of distributed memory message passing solutions.

Therefore, 3Leaf network shared memory drastically modifies the paradigm for applications designed for cluster computing. With network shared memory, the cluster applications are freed from the overhead and latencies of message passing and distributed synchronization. The support for hardware read-modify-write operations in network shared memory enables very scalable and non-blocking synchronization methods formerly available only in traditional shared memory platforms. Furthermore, the memory devoted to the cluster computing environment in this way is utilized much more effectively. There is no longer the need to have multiple copies of the same data, along with the associated messaging overhead to replicate the data between nodes. The lack of memory duplication leads to increasing a cluster's flexibility while reducing its cost.

Development Flow for HPC

The raw runtime for an HPC application only takes into account part of the solution cost. The development of HPC applications on cluster computers also needs careful consideration. For some HPC applications, development time is a significant time and cost component of the HPC problem. Broadly speaking, there are two paradigms to consider - parallel and pipelined.

In the parallel model, a computation problem is broken into identical or similar, computational threads. The data are then divided among the computing threads. These problems are often referred to as embarrassingly parallelizable, but there are also problems of this kind that have a high communication to computation ratio and do not run well on message passing clusters. An example of this kind of application is the simulation of fluid flows.

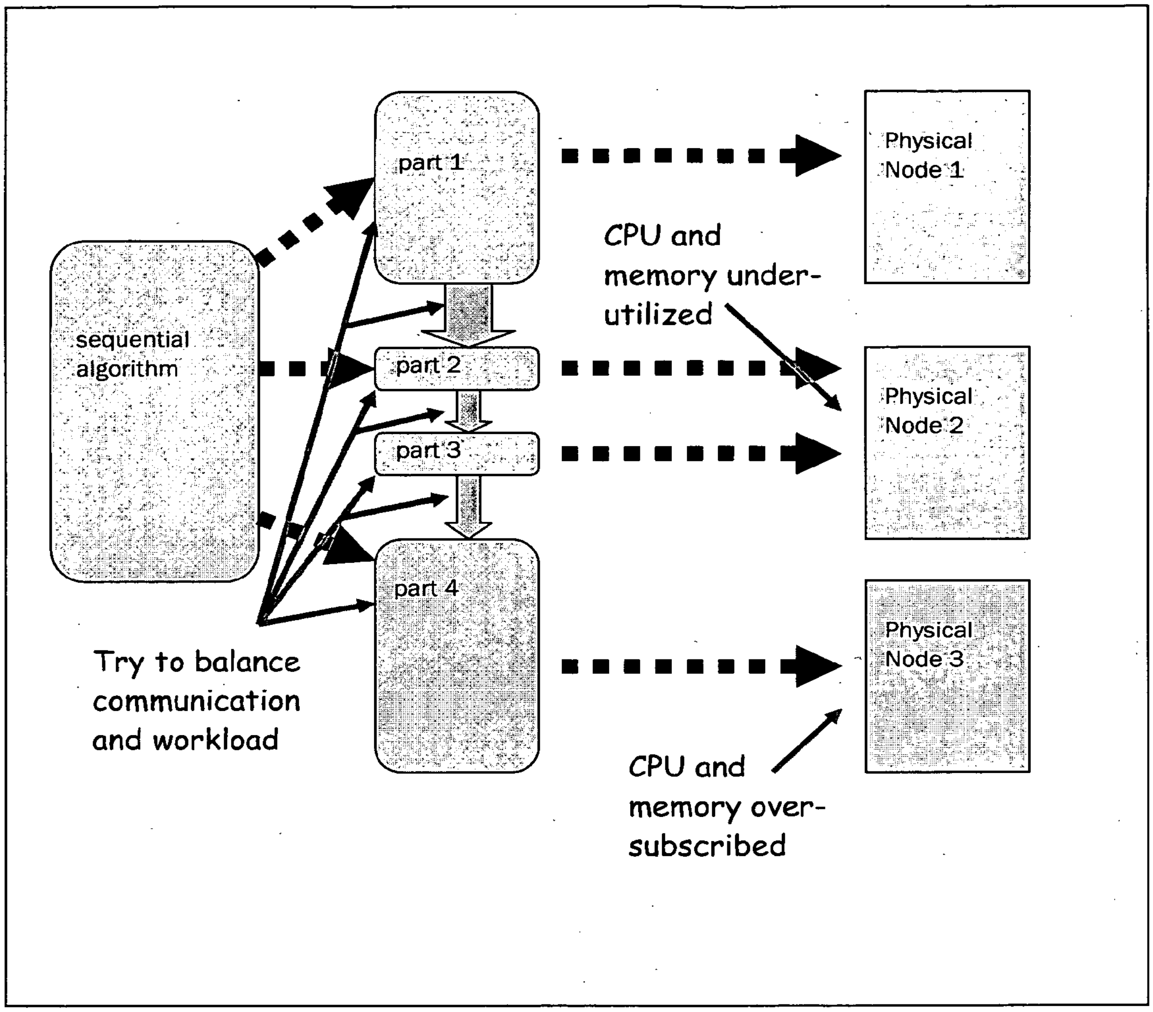
Another method of parallelizing an application is to divide the task into sequential subtasks that can map onto a computation pipeline. An example application that maps onto a pipeline nicely is image processing, where a series of image transformations can be chained together in a sequential pipeline.

The challenge for the HPC programmer is to map the problem into pipeline stages that satisfy three constraints:

1. minimize communication between stages
2. map the number of pipeline stages to match the number of computational nodes available
3. size each of the pipeline stages to fit within available CPU and memory resources of each node (optimally use all the available resources of each node)

3Leaf's DDC removes these constraints from the programmer by virtualizing CPU and memory resources on a cluster based computer. ***This allows programmers to spend more time focusing on solving the problem instead of mapping the problem onto a given machine configuration***. Different CPU and memory configurations can be created to try different problem partitions, enabling application optimization without regard to the physical hardware constraints.

Figure 1 Mapping a Sequential Problem onto a Compute Pipeline


part 1
Physical Node 1
CPU and memory under-utilized
sequential algorithm
part 2
Physical Node 2
part 3
part 4
Physical Node 3
Try to balance communication and workload
CPU and memory over-subscribed

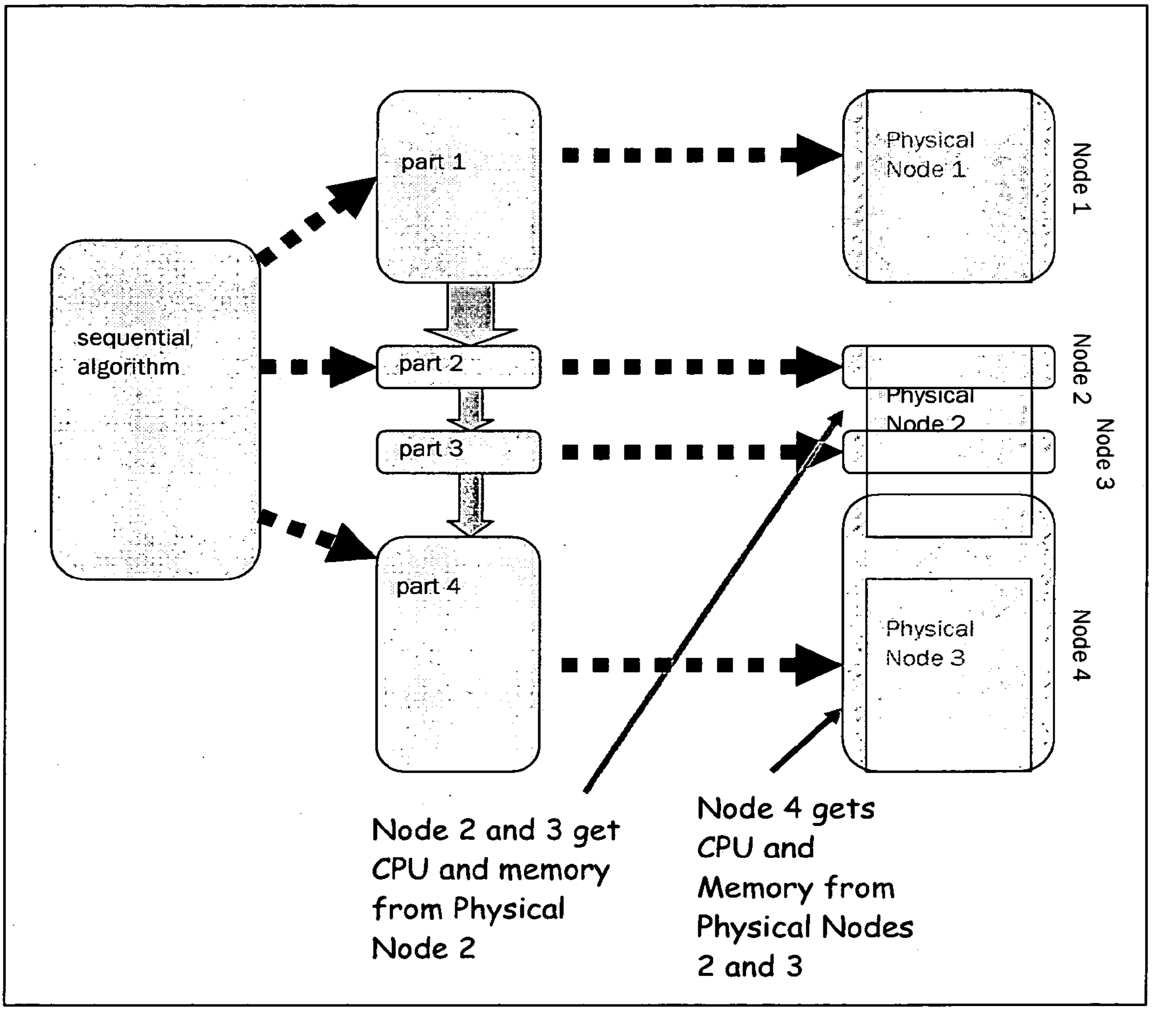
In a 3Leaf DDC, communication bandwidth and latency are reduced by allowing memory to be shared between nodes. The multi-node coherent memory allows for shared memory implementations of message passing (through a library like OpenMPI) or through shared memory exposed to the application. Thus, programmers can consider partitions that would have been impractical on traditional HPC clusters due to bandwidth and latency limitations of non-shared memory message passing.

Since 3Leaf implements Virtual Compute Nodes, the cores on individual servers can be grouped in a way that matches the number of compute nodes required by the application pipeline. For example, consider a physical machine consisting of 3 physical servers. If the problem better matches a machine that has 4 compute nodes, the CPUs on the three individual servers can be grouped into 4 logical servers. Indeed, one of the prime advantages of cluster computing over Scale-Up computing is that the cluster can be sized to match the problem. The 3Leaf Dynamic Data Center takes this notion even further, by allowing hardware clusters to be partitioned and allocated to fit each and every application.

Traditional HPC clusters have hard constraints on the number of CPU cores and amount of physical memory available. Each server has only a fixed amount of memory and CPU core resources. This forces the programmer to consider balancing the CPU and memory resources of each stage of the pipeline with available hardware resources. With the 3Leaf Dynamic Data Center, each computational node can be sized with the number of compute and amount of memory resources required.

Finally, the constraints on memory in traditional HPC clusters may lead some users or IT professionals to provision each node with the maximum amount of memory so that the programmer's burden is mitigated and system flexibility is increased. This solution results in wasted resources and results in higher capital and operating expenses. The 3Leaf Dynamic Data Center network shared memory solves this problem with its cluster wide shared memory, resulting in capital and operating expense efficiencies.

Figure 2 - 3Leaf Relieves Hardware Partition Limitations



Building on Industry Trends

Multiple industry trends have contributed to enable the 3Leaf solution for compute and memory virtualization:

Advanced commodity 64 bit processors

Today's processors include features that meet the needs of enterprise class data-centers, with seamless support for virtualization, 64 bit addressing, memory fault detection and recovery, hardware enforced secure portioning, multi-core going from 4 today to 6 and 8 tomorrow, and rapidly increasing cache sizes.

Network switches delivering increasing bandwidth with decreasing latency

As silicon process technology evolves, latencies for commodity switch chips continue to fall. Not only has the bandwidth of single chip switches reached Terabits per second, the time it takes from a signal arriving at the input pin to appear at the output pin is approaching the same order of magnitude as DRAM access times, with IB switches approaching switching times of 100nS and Ethernet switches in hot pursuit.

Enterprise Capable commodity operating systems

Commodity operating systems such as Windows and Linux now provide ccNUMA optimizations, and also support the hot plug and hot unplug of devices, CPU and memory. In addition, both Operating Systems and today's enterprise applications are fully compatible with virtualized servers (via a hypervisor)

3leaf builds on these trends

The Coherent NIC from 3Leaf lets threads running on cores in one server interact directly with threads and memory located on another server, and is enabled by the bandwidth and latency of modern switches. The Distributed Machine Monitor from 3Leaf insulates an Operating System from a distributed and changing set of core and memory resources, and is enabled by the advanced features of today's processors and enterprise class Windows and Linux operating systems.

Conclusion

HPC computing continues its migration towards cluster based solutions for flexibility, cost and performance reasons. However, traditional clusters do not provide the rich and critical performance features of coherent shared memory. 3Leaf's Dynamic Data Center brings the advantages of shared memory supercomputers to cluster based computers.

3Leaf is enabling the Dynamic Data Center with next generation server virtualization that addresses today's changing business needs by providing the on-demand resources and flexibility that can literally revolutionize operational and development efficiencies. Virtualization of CPU, memory, and I/O resources enables the creation of a pool of server resources that can span across multiple physical machines and be allocated or de-allocated as needed. Coherent shared memory among commodity x86 compute nodes provides a high performance compute cluster for HPC applications. Easy repurposing and migration allows machines to remain fully utilized as workloads change during the course of the day, week, or month. Fast and flexible provisioning allows machines to be mass deployed within a heterogeneous environment, drastically reducing the time and cost of developing applications and provisioning new servers and the applications they support.

3Leaf's Dynamic Data Center is truly dynamic, nimbly responding to the changing requirements of HPC applications, while drastically reducing both capital and operating expenses.